行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　　　　　　　　　　　　　　　　　　　　　　　※

※　　　　設限資料參數模型之適合度檢定　　　　※

※　　　　　　　　　　　　　　　　　　　　　　　　　※

※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC 90－2118－M－032－007

執行期間：90 年 08 月 01 日至 91 年 07 月 31 日

計畫主持人：黃連成

共同主持人：

計畫參與人員：楊鎮魁、吳永通、蔡英洲

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：淡江大學數學系

中　華　民　國　91 年 8 月 1 日

行政院國家科學委員會專題研究計畫成果報告

Preparation of NSC Project Reports

計畫編號:NSC 90-2118-M-032-007

執行期限:90年08月01日至91年07月31日

主持人: 黃連成

計畫參與人員: 楊鎮魁, 吳永通, 蔡英洲

執行機構及單位名稱: 淡江大學數學系

A Note on Goodness-of-Fit Tests for Censored Data

Leng-Cheng Hwang

Department of Mathematics

Tamkang University

Tamsui, Taipei, Taiwan, R.O.C.

Abstract

Goodness-of-fit tests for censored data usually including parameters estimated after experimental termination were established in many papers. It is shown that a proposed process with the maximum likelihood estimate of the parameters, depending on the calendar time, in parametric model of censored data with staggered entry is asymptotically Gaussian martingale in this paper. Some goodness-of-fit tests are made by the weak convergence of the proposed process.

AMS 1991 subject classification: 62F03, 62L10.

Key words and phrases: Censored data, goodness of fit, staggered entry, weak convergence.

1. Introduction

Censored survival data appears in the areas of biostatistics, reliability, life history data and actuarial statistics, etc. Some attention has been given to the problem of model checking for censored data, for example, Arjas (1988), Lin and Wei (1991) and McKeague and Utikal (1991) for Cox regression model, and Akritas (1988) and Hjort (1990) for parametric model. Goodness-of-fit tests in all these discussions are for the situation that the parameters are estimated after experimental termination. Here parametric model of censored data with staggered entry is considered, some goodness-of-fit tests are proposed in it. The parameters estimated in these tests depend on the calendar time.

Let (Y_j, Z_j, X_j, C_j) be a sequence of independent and identically distributed random vectors with Y_j denoting the entry time, Z_j the covariate, X_j the survival time and C_j the censoring time of the ith subject in a clinical trial. Assume that Y_j and X_j are independent; conditional on Y_j and Z_j, X_j and C_j are independent. Let the hazard function of X_j given Z_j be of the form

$$\lambda(\cdot, Z_j, \theta) \tag{1.1}$$

for some $\theta \in R^p$, and assume that $\lambda(t, Z_j, \theta)$ has continuous third partial derivatives in θ .

The data available at calendar time t is

$$\{Y_j \wedge s, Z_j, (Y_j + X_j \wedge C_j) \wedge s, 1_{[(Y_j+X_j)\wedge s \leq (Y_j+C_j)\wedge s]} | s \leq t, j = 1, \cdots, n\}.$$

Let $\mathcal{F}_t^n$ be σ-field generated by the data at time t, and

$$M_j(t, \theta) = 1_{[Y_j+X_j,\infty)}(t \wedge (Y_j + C_j)) - \int_0^t \lambda(s - Y_j, Z_j, \theta) 1_{(Y_j, Y_j+X_j\wedge C_j]}(s) ds.$$

We note that $M_j(t, \theta)$ is a $\mathcal{F}_t^n$-martingale. According to Chang and Hsiung (1998),

the log-likelihood of the data at time t is

$$L_n(t,\theta) = \sum_{j=1}^{n} \int_0^t log\lambda(u, Z_j, \theta) dN_j{}^t(u) - \sum_{j=1}^{n} \int_0^t \lambda(u, Z_j, \theta) H_j{}^t(u) du,$$

where $N_j{}^t(u) = 1_{[X_j,\infty)}(u \wedge C_j \wedge (t - Y_j)^+)$ and $H_j{}^t(u) = 1_{(0, X_j \wedge C_j \wedge (t-Y_j)^+]}(u)$, and then the likelihood score process

$$\begin{aligned} U_{n,l}(t,\theta) &= \frac{\partial}{\partial \theta_l} L_n(t,\theta) \\ &= \sum_{j=1}^{n} \int_0^t \frac{\frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta)}{\lambda(u - Y_j, Z_j, \theta)} 1_{(Y_j,\infty)}(u) dM_j(u,\infty) \end{aligned}$$

is a $\mathcal{F}_t{}^n$-martingale for every $l = 1, \cdots, p$.

Let $U_n(t,\theta) = (U_{n,1}(t,\theta), \cdots, U_{n,p}(t,\theta))$, and a solution $\hat{\theta}_n(t)$ of $U_n(t, \hat{\theta}_n(t)) = 0$ be a maximum likelihood estimate of θ at calendar time t. We consider the stochastic process

$$H_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} M_j(t, \hat{\theta}_n(t)).$$

The asymptotic properties of the process are studied, and then some goodness-of-fit tests are made in the paper.

2. Main results

Let E denote the range of Z_1 and Θ be a neighborhood of θ_0, where θ_0 be the true parameter. In the remainder of this paper we assume for any $t_0 > 0$, that $\lambda(\cdot,\cdot,\cdot)$ is bounded away from 0 and that the first, second and third partial derivatives of $\lambda(\cdot,\cdot,\cdot)$ in θ are bounded on $(0, t_0] \times E \times \Theta$.

For convenience, let

$$\begin{aligned} \phi_{p+1,p+1}(t,\theta_0) &= E\{\lambda(t - Y_1, Z_1, \theta_0) 1_{(Y_1, Y_1 + X_1 \wedge C_1]}(t)\} \\ \phi_{l,p+1}(t,\theta_0) &= \phi_{p+1,l}(t,\theta_0) = E\{\frac{\partial \lambda}{\partial \theta_l}(t - Y_1, Z_1, \theta_0) 1_{(Y_1, Y_1 + X_1 \wedge C_1]}(t)\} \\ \phi_{l,k}(t,\theta_0) &= E\{\frac{\frac{\partial \lambda}{\partial \theta_l}(t - Y_1, Z_1, \theta_0) \frac{\partial \lambda}{\partial \theta_k}(t - Y_1, Z_1, \theta_0)}{\lambda(t - Y_1, Z_1, \theta_0)} 1_{(Y_1, Y_1 + X_1 \wedge C_1]}(t)\}, \end{aligned}$$

where $l, k = 1, \cdots, p$, and $\triangle_{1l}(t, \theta_0) = \int_0^t \phi_{l,p+1}(u, \theta_0)du$ and $\triangle_{2lk}(t, \theta_0) = \int_0^t \phi_{l,k}(u, \theta_0)du$ be the l-th and (l, k)-th entry of $1 \times p$ matrix $\triangle_1(t, \theta_0)$ and $p \times p$ matrix $\triangle_2(t, \theta_0)$, respectively.

Theorem 1. Under the assumed model (1.1), if $\triangle_2(a, \theta_0)$ is invertible for some $a > 0$, then $H_n(t)$ converges weakly on $D[a, \infty)$ to a continuous Gaussian martingale with variance $g(t, \theta_0)$, where $g(t, \theta_0) = -\triangle_1(t, \theta_0)\triangle_2{}^{-1}(t, \theta_0)\triangle_1{}'(t, \theta_0) + \int_0^t \phi_{p+1,p+1}(u, \theta_0)du$.

This theorem tells us that the model should be rejected if H_n is significantly different from zero, as measured by some suitable functional. Because that the variance $g(t, \theta_0)$ depends on the unknown parameter θ_0, we need to give an estimator of $g(t, \theta_0)$ in order to establish some goodness-of-fit tests based on it. Let

$$\begin{aligned}
\hat{\triangle}_{1l}(t, \hat{\theta}_n(t)) &= \frac{1}{n}\sum_{j=1}^{n} \int_0^t \frac{\frac{\partial\lambda}{\partial\theta_l}(u - Y_j, Z_j, \hat{\theta}_n(t))}{\lambda(u - Y_j, Z_j, \hat{\theta}_n(t))} d1_{[Y_j+X_j,\infty)}(u \wedge (Y_j + C_j)) \\
\hat{\triangle}_{2lk}(t, \hat{\theta}_n(t)) &= \frac{1}{n}\sum_{j=1}^{n} \int_0^t \frac{\frac{\partial\lambda}{\partial\theta_l}(u - Y_j, Z_j, \hat{\theta}_n(t))\frac{\partial\lambda}{\partial\theta_k}(u - Y_j, Z_j, \hat{\theta}_n(t))}{\lambda^2(u - Y_j, Z_j, \hat{\theta}_n(t))} \cdot \\
&\quad d1_{[Y_j+X_j,\infty)}(u \wedge (Y_j + C_j))
\end{aligned}$$

for every $l, k = 1, \cdots, p$, and $\hat{\triangle}_{1l}(t, \hat{\theta}_n(t))$ and $\hat{\triangle}_{2lk}(t, \hat{\theta}_n(t))$ be the l-th and (l, k)-th entry of $1 \times p$ matrix $\hat{\triangle}_1(t, \hat{\theta}_n(t))$ and $p \times p$ matrix $\hat{\triangle}_2(t, \hat{\theta}_n(t))$, respectively.

Theorem 2. Under the same conditions for Theorem 1, $\hat{\triangle}_{1l}(t, \hat{\theta}_n(t))$, $\hat{\triangle}_{2lk}(t, \hat{\theta}_n(t))$ and $\frac{1}{n}\sum_{j=1}^{n} \int_0^t d1_{[Y_j+X_j,\infty)}(u \wedge (Y_j + C_j))$ are consistent for estimating $\triangle_{1l}(t, \theta_0)$, $\triangle_{2lk}(t, \theta_0)$ and $\int_0^t \phi_{p+1,p+1}(u, \theta_0)du$ for any $t \geq a$, respectively.

Let

$$\hat{g}(t, \hat{\theta}_n(t)) = -\hat{\triangle}_1(t, \hat{\theta}_n(t))\hat{\triangle}_2^{-1}(t, \hat{\theta}_n(t))\hat{\triangle}_1'(t, \hat{\theta}_n(t)) + \frac{1}{n}\sum_{j=1}^{n} \int_0^t d1_{[Y_j+X_j,\infty)}(u \wedge (Y_j + C_j)),$$

which is consistent estimator of $g(t, \theta_0)$, and we define the test statistic

$$T_m = \sum_{i=1}^{m} \left(\frac{H_n(t_i) - H_n(t_{i-1})}{\sqrt{\hat{g}(t_i, \hat{\theta}_n(t_i)) - \hat{g}(t_{i-1}, \hat{\theta}_n(t_{i-1}))}}\right)^2,$$

where $0 = t_0 < a \leq t_1 < t_2 < \cdots < t_m < \infty$. It is clear that the limit distribution of T_m is χ^2 distributed with m degrees of freedom.

3. Proofs

We will now develop some auxiliary results and apply these results to prove Theorem 1 and Theorem 2. For convenience, the conditions on the following lemmas is assumed as the same as Theorem 1, and omitted in here.

Lemma 1. The process $\sqrt{n}(\hat{\theta}_n(t) - \theta_0)$ shares the same weak limit on $[a, \infty)$ as $\frac{1}{\sqrt{n}} U_n(t, \theta_0) \triangle_2{}^{-1}(t, \theta_0)$.

Proof. The lemma follows from (4.4), (4.5) and Theorem 3.1 in Chang and Hsiung (1988).

Lemma 2. For every $t_0 > 0$ and $l = 1, \cdots, p$, we have

$$sup_{t\in[a,t_0]} |\frac{1}{n}\sum_{j=1}^{n} \int_0^t \frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta_n) 1_{(Y_j, Y_j + X_j \wedge C_j]}(u) du - \triangle_{1l}(t, \theta_0)| \xrightarrow{P} 0$$

where θ_n lies between θ_0 and $\hat{\theta}_n(t)$.

Proof. Write

$$\begin{aligned}
& sup_{t\in[a,t_0]} |\frac{1}{n}\sum_{j=1}^{n} \int_0^t \frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta_n) 1_{(Y_j, Y_j + X_j \wedge C_j]}(u) du - \triangle_{1l}(t, \theta_0)| \\
\leq\ & sup_{t\in[a,t_0]} |\frac{1}{n}\sum_{j=1}^{n} \int_0^t (\frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta_n) - \frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta_0)) 1_{(Y_j, Y_j + X_j \wedge C_j]}(u) du| \\
& + sup_{t\in[a,t_0]} |\frac{1}{n}\sum_{j=1}^{n} \int_0^t \frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta_0) 1_{(Y_j, Y_j + X_j \wedge C_j]}(u) du - \triangle_{1l}(t, \theta_0)| \\
=\ & I_a + I_b,
\end{aligned}$$

say. Using the bounded convergence theorem,

$$\begin{aligned}
I_b \ &\leq \ \int_0^{t_0} |\frac{1}{n}\sum_{j=1}^{n} \frac{\partial \lambda}{\partial \theta_l}(u - Y_j, Z_j, \theta_0) 1_{(Y_j, Y_j + X_j \wedge C_j]}(u) du - \phi_{l,p+1}(u, \theta_0)| du \\
&\to \ 0 \text{ a.s.}
\end{aligned}$$

Applying the Mean-Value Theorem, we get

$$I_a = sup_{t\in[a,t_0]}|\frac{1}{n}\sum_{j=1}^{n}\int_0^t \sum_{k=1}^{p}\frac{\partial^2\lambda}{\partial\theta_k\partial\theta_l}(u-Y_j,Z_j,\tilde{\theta}_n)(\tilde{\theta}_n-\theta_0)_k 1_{(Y_j,Y_j+X_j\wedge C_j]}(u)du|$$

for some $\tilde{\theta}_n$ lying between θ_0 and θ_n, where $(\tilde{\theta}_n-\theta_0)_k$ denotes the k-th component of $\tilde{\theta}_n-\theta_0$. Using theorem 4.1 of Chang and Hsiung (1988), for any $r>0$ and there exists $\epsilon>0$ depending on Θ,

$$\begin{aligned}
P\{I_a>r\} &= P\{I_a>r, sup_{t\in[a,t_0]}|\hat{\theta}_n(t)-\theta_0|\le\epsilon\} \\
&\quad +P\{I_a>r, sup_{t\in[a,t_0]}|\hat{\theta}_n(t)-\theta_0|>\epsilon\} \\
&\le P\{k\ sup_{t\in[a,t_0]}|\tilde{\theta}_n-\theta_0|>r\}+P\{sup_{t\in[a,t_0]}|\hat{\theta}_n(t)-\theta_0|>\epsilon\} \\
&\to 0,
\end{aligned}$$

for some constant $k>0$. This completes the proof of the lemma.

Lemma 3. $W_n(t,\theta_0)=\frac{1}{\sqrt{n}}(U_n(t,\theta_0),\sum_{j=1}^n\int_0^t dM_j(u,\theta_0))$ converges weakly on $[0,\infty)$ to a continuous Gaussian martingale with the (l,k)-th component of covariance $\int_0^t\phi_{l,k}(u,\theta_0)du$ for every $l,k=1,\cdots,p+1$.

Proof. It follows from the strong law of large numbers and Fubini's theorem that the predictable covariation process

$$\begin{aligned}
&\langle\frac{1}{\sqrt{n}}U_{n,l}(\cdot,\theta_0),\frac{1}{\sqrt{n}}U_{n,k}(\cdot,\theta_0)\rangle_t \\
&= \frac{1}{n}\sum_{j=1}^n\int_0^t\frac{\frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\theta_0)\frac{\partial\lambda}{\partial\theta_k}(u-Y_j,Z_j,\theta_0)}{\lambda(u-Y_j,Z_j,\theta_0)}1_{(Y_j,Y_j+X_j\wedge C_j]}(u)du \\
&\to \int_0^t\phi_{l,k}(u,\theta_0)du \ \text{ a.s.},
\end{aligned}$$

and the ϵ-jump part of $\frac{1}{\sqrt{n}}U_{n,l}(\cdot,\theta_0)$

$$\begin{aligned}
\langle\frac{1}{\sqrt{n}}U_{n,l\epsilon}(\cdot,\theta_0)\rangle_t &= \frac{1}{n}\sum_{j=1}^n\int_0^t\frac{(\frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\theta_0))^2}{\lambda(u-Y_j,Z_j,\theta_0)}1_{(Y_j,Y_j+X_j\wedge C_j]}(u)\cdot \\
&\qquad 1_{\{|\frac{1}{\sqrt{n}}\frac{\frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\theta_0)}{\lambda(u-Y_j,Z_j,\theta_0)}1_{(Y_j,\infty)}(u)|>\epsilon\}}du \\
&\to 0 \ \text{ a.s.}
\end{aligned}$$

Using the similar arguments for other components in W_n, hence by the martingale central limit theorem, the lemma follows.

Proof of Theorem 1.

The Mean-Value Theorem gives

$$\begin{aligned} & H_n(t) \\ = & \frac{1}{\sqrt{n}}\sum_{j=1}^{n} M_j(t,\theta_0) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^t (\lambda(s-Y_j,Z_j,\hat{\theta}_n(t)) - \lambda(s-Y_j,Z_j,\theta_0))\cdot \\ & 1_{(Y_j,Y_j+X_j\wedge C_j]}(s)ds \\ = & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^t dM_j(s,\theta_0) - \sqrt{n}(\hat{\theta}_n(t)-\theta_0)\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^t \nabla\lambda(s-Y_j,Z_j,\tilde{\theta}_n(t))\cdot \\ & 1_{(Y_j,Y_j+X_j\wedge C_j]}(s)ds \end{aligned}$$

for some $\tilde{\theta}_n(t)$ lying between θ_0 and $\hat{\theta}_n(t)$ and $\nabla\lambda(s-Y_j,Z_j,\theta) = (\frac{\partial\lambda}{\partial\theta_1}(s-Y_j,Z_j,\theta),\cdots,$ $\frac{\partial\lambda}{\partial\theta_p}(s-Y_j,Z_j,\theta))'$. Then, using Lemma 1 and Lemma 2, we get that $H_n(t)$ is asymptotically equivalent to $W_n(t)(-\triangle_1(t,\theta_0)\triangle_2{}^{-1}(t,\theta_0),1)'$. Hence, by Lemma 3 and Theorem 5.1 in Billingsley (1968), $H_n(t)$ converges weakly to a continuous Gaussian martingale with variance $g(t,\theta_0)$.

Proof of Theorem 2.

One notes that

$$\begin{aligned} & \hat{\triangle}_{1l}(t,\hat{\theta}_n(t)) - \triangle_{1l}(t,\theta_0) \\ = & \{\frac{1}{n}\sum_{j=1}^{n}\int_0^t (\frac{\frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\hat{\theta}_n(t))}{\lambda(u-Y_j,Z_j,\hat{\theta}_n(t))} - \frac{\frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\theta_0)}{\lambda(u-Y_j,Z_j,\theta_0)})d1_{[Y_j+X_j,\infty)}(u\wedge(Y_j+C_j))\} \\ & +\{\frac{1}{n}\sum_{j=1}^{n}\int_0^t \frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\theta_0)1_{(Y_j,Y_j+X_j\wedge C_j]}(u)du \\ & \quad - E\int_0^t \frac{\partial\lambda}{\partial\theta_l}(u-Y_1,Z_1,\theta_0)1_{(Y_1,Y_1+X_1\wedge C_1]}(u)du\} \\ & +\{\frac{1}{n}\sum_{j=1}^{n}\int_0^t \frac{\frac{\partial\lambda}{\partial\theta_l}(u-Y_j,Z_j,\theta_0)}{\lambda(u-Y_j,Z_j,\theta_0)}dM_j(u,\theta_0)\} \\ = & II_a + II_b + II_c, \end{aligned}$$

say. We can easily get that II_b, $II_c \to 0$ a.s. and $II_a \xrightarrow{P} 0$ by using analogous arguments as I_a in the proof of Lemma 2.

The proof of other statements in theorem is the same as this. The theorem follows.

References

Akritas, M.G., 1988. Pearson-type goodness-of-fit tests: the univariate case, J. Amer. Statist. Assoc. **83**, 222-230.

Arjas, E., 1988. A graphical method for assessing goodness of fit in Cox's proportional hazards model, J. Amer. Statist. Assoc. **83**, 204-212.

Billingsley, P., 1968. Convergence of Probability Measures, Wiley, New York.

Chang, I.-S. and Hsiung, C.A., 1988. Likelihood process in parametric model of censored data with staggered entry-Asymptotic properties and applications, J. Multivariate Anal. **24**, 31-45.

Hjort, N.L., 1990. Goodness of fit tests in models for life history data based on cumulative hazard rates, Ann. Statist. **18**, 1221-1258.

Lin, D.Y. and Wei, L.J., 1991. Goodness-of-fit tests for the general Cox regression model, Statist. Sinica **1**, 1-17.

McKeague, I.W. and Utikal, K.J., 1991. Goodness-of-fit tests for additive hazards and proportional hazards models. Scand. J. Statist. **18**, 177-195.